Differential Privacy and the 2020 Decennial Census

Michael Hawes
Senior Advisor for Data Access and Privacy
Research and Methodology Directorate
U.S. Census Bureau

Puget Sound Regional Council
June 2, 2020

Shape
your future
START HERE >

United States®
Census
2020

Acknowledgements

This presentation includes work by the Census Bureau's 2020 Disclosure Avoidance System development team, Census Bureau colleagues, and our collaborators, including: John Abowd, Tammy Adams, Robert Ashmead, Craig Corl, Ryan Cummings, Jason Devine, John Fattaleh, Simson Garfinkel, Nathan Goldschlag, Michael Hawes, Michael Hay, Cynthia Hollingsworth, Michael Ikeda, Kyle Irimata, Dan Kifer, Philip Leclerc, Ashwin Machanavajjhala, Christian Martindale, Gerome Miklau, Claudia Molinar, Brett Moran, Ned Porter, Sarah Powazek, Vikram Rao, Chris Rivers, Anne Ross, Ian Schmutte, William Sexton, Rob Sienkiewicz, Matthew Spence, Tori Velkoff, Lars Vilhuber, Bei Wang, Tommy Wright, Bill Yates, and Pavel Zhurlev.

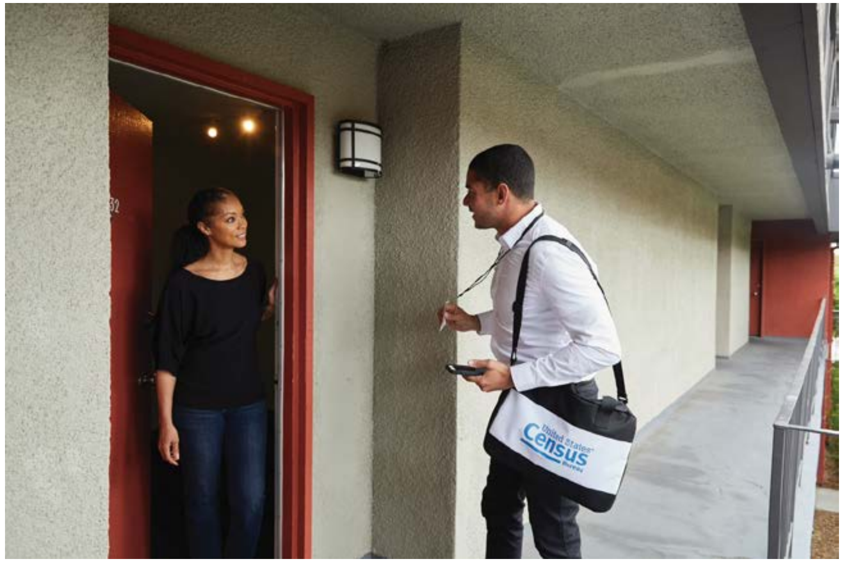
For more information and technical details relating to the issues discussed in these slides, please contact the author at michael.b.hawes@census.gov.

Any opinions and viewpoints expressed in this presentation are the author's own, and do not represent the opinions or viewpoints of the U.S. Census Bureau.

Our Commitment to Data Stewardship

Data stewardship is central to the Census Bureau's mission to produce high-quality statistics about the people and economy of the United States.

Our commitment to protect the privacy of our respondents and the confidentiality of their data is both a legal obligation and a core component of our institutional culture.



It’s the Law

> *“To stimulate public cooperation necessary for an accurate census…Congress has provided assurances that information furnished by individuals is to be treated as confidential. Title 13 U.S.C. §§ 8(b) and 9(a) explicitly provide for nondisclosure of certain census data, and **no discretion is provided to the Census Bureau on whether or not to disclose such data**…”* (U.S. Supreme Court, Baldrige v. Shapiro, 1982)

Title 13, Section 9 of the United State Code prohibits the Census Bureau from releasing identifiable data “furnished by any particular establishment or individual.”

Census Bureau employees are sworn for life to safeguard respondents’ information.

Penalties for violating these protections can include fines of up to $250,000, and/or imprisonment for up to five years!

Keeping the Public’s Trust

Safeguarding the public’s data is about more than just complying with the law!

The quality and accuracy of our censuses and surveys depend on our ability to keep the public’s trust.

In an era of declining trust in government, increasingly common corporate data breaches, and declining response rates to surveys, we must do everything we can to keep our promise to protect the confidentiality of our respondent’s data.

Upholding our Promise: Today and Tomorrow

We cannot merely consider privacy threats that exist today.

We must ensure that our disclosure avoidance methods are also sufficient to protect against the threats of tomorrow!

The Privacy Challenge

Every time you release any statistic calculated from a confidential data source you “leak” a small amount of private information.

If you release too many statistics, too accurately, you will eventually reveal the entire underlying confidential data source.

Dinur, Irit and Kobbi Nissim (2003) “Revealing Information while Preserving Privacy” PODS, June 9-12, 2003, San Diego, CA

The Growing Privacy Threat

More Data and Faster Computers!

In today’s digital age, there has been a proliferation of databases that could potentially be used to attempt to undermine the privacy protections of our statistical data products.

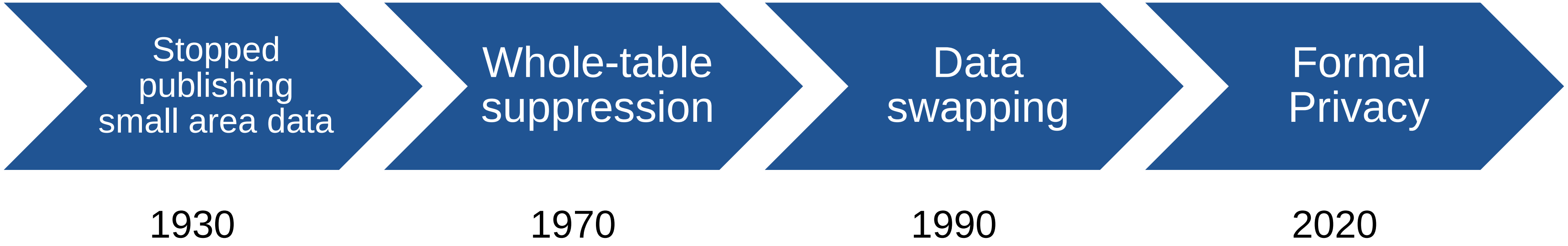
Similarly, today’s computers are able to perform complex, large-scale calculations with increasing ease.

These parallel trends represent new threats to our ability to safeguard respondents’ data.

The Census Bureau’s Privacy Protections Over Time

Throughout its history, the Census Bureau has been at the forefront of the design and implementation of statistical methods to safeguard respondent data.

Over the decades, as we have increased the number and detail of the data products we release, so too have we improved the statistical techniques we use to protect those data.

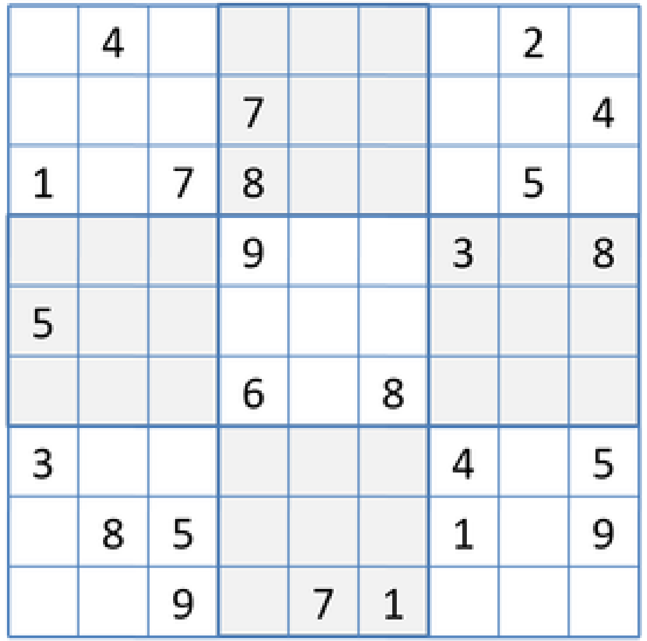


Reconstruction

The recreation of individual-level data from tabular or aggregate data.

If you release enough tables or statistics, eventually there will be a unique solution for what the underlying individual-level data were.

Computer algorithms can do this very easily.



	4						2	
			7					4
1		7	8				5	
			9			3		8
5								
			6		8			
3						4		5
	8	5				1		9
		9		7	1			

Reconstruction: An Example

	Count	Median Age	Mean Age
Total	7	30	38
Female	4	30	33.5
Male	3	30	44
Black	4	51	48.5
White	3	24	24
Married	4	51	54
Black Female	3	36	36.7

Reconstruction: An Example

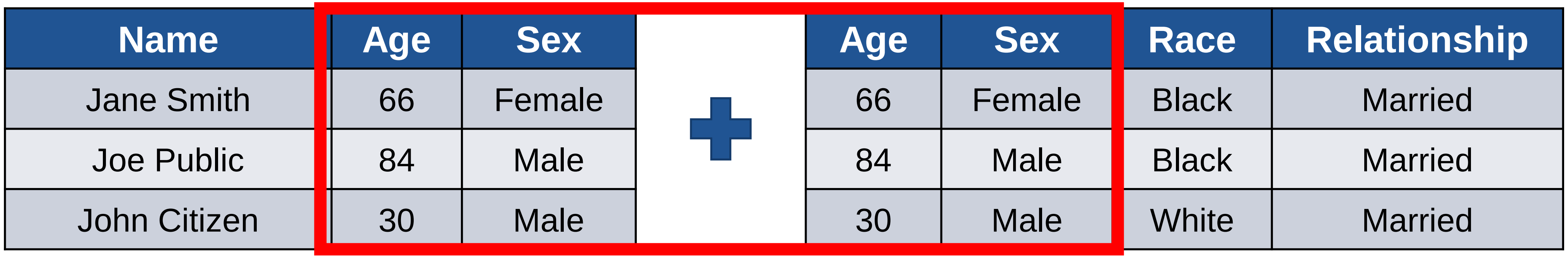
	Count	Median Age	Mean Age
Total	7	30	38
Female	4	30	33.5
Male	3	30	44
Black	4	51	48.5
White	3	24	24
Married	4	51	54
Black Female	3	36	36.7

Age	Sex	Race	Relationship
66	Female	Black	Married
84	Male	Black	Married
30	Male	White	Married
36	Female	Black	Married
8	Female	Black	Single
18	Male	White	Single
24	Female	White	Single

This table can be expressed by 164 equations. Solving those equations takes 0.2 seconds on a 2013 MacBook Pro.

Re-identification

Linking public data to external data sources to re-identify specific individuals within the data.



Name	Age	Sex
Jane Smith	66	Female
Joe Public	84	Male
John Citizen	30	Male

External Data

➕

Age	Sex	Race	Relationship
66	Female	Black	Married
84	Male	Black	Married
30	Male	White	Married

Confidential Data

In the News

Reconstruction and Re-identification are not just theoretical possibilities…they are happening!

- **Massachusetts Governor's Medical Records (Sweeney, 1997)**
- **AOL Search Queries (Barbaro and Zeller, 2006)**
- **Netflix Prize (Narayanan and Shmatikov, 2008)**
- **Washington State Medical Records (Sweeney, 2015)**
- **and many more…**

Reconstructing the 2010 Census

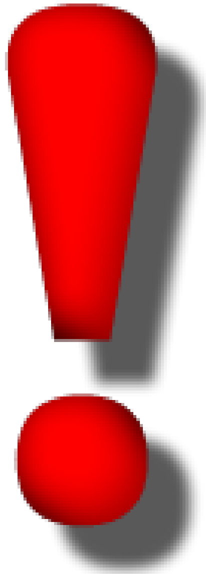
- The 2010 Census collected information on the age, sex, race, ethnicity, and relationship (to householder) status for ~309 Million individuals. (1.9 Billion confidential data points)
- The 2010 Census data products released over 150 billion statistics
- We conducted an internal experiment to see if we could reconstruct and re-identify the 2010 Census records.

Reconstructing the 2010 Census: What Did We Find?

1. On the 309 million reconstructed records, census block and voting age (18+) were correctly reconstructed for all records and for all 6,207,027 inhabited blocks.
2. Block, sex, age (in years), race (OMB 63 categories), and ethnicity were reconstructed:
    1. Exactly for 46% of the population (142 million individuals)
    2. Within +/- one year for 71% of the population (219 million individuals)
3. Block, sex, and age were then linked to commercial data, which provided putative re-identification of 45% of the population (138 million individuals).
4. Name, block, sex, age, race, ethnicity were then compared to the confidential data, which yielded confirmed re-identifications for 38% of the putative re-identifications (52 million individuals).
5. For the confirmed re-identifications, race and ethnicity are learned correctly, though the attacker may still have uncertainty.

The Census Bureau’s Decision

- Advances in computing power and the availability of external data sources make database reconstruction and re-identification increasingly likely.
- The Census Bureau recognized that its traditional disclosure avoidance methods are increasingly insufficient to counter these risks.
- To meet its continuing obligations to safeguard respondent information, the Census Bureau has committed to modernizing its approach to privacy protections.



Differential Privacy

aka “Formal Privacy”

-quantifies the precise amount of privacy risk…

-for all calculations/tables/data products produced…

-no matter what external data is available…

-now, or at any point in the future!

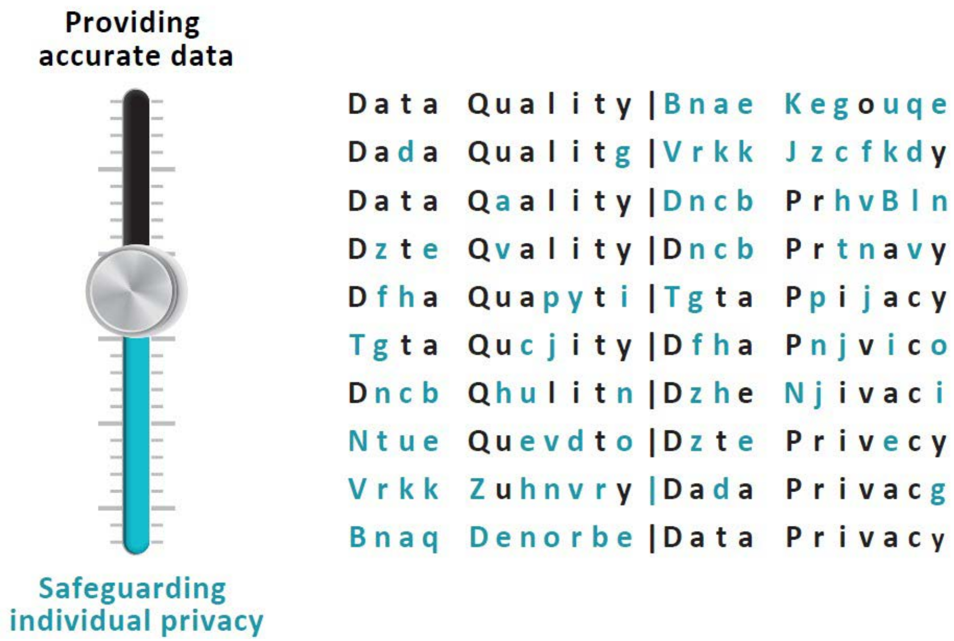
Precise amounts of noise

Differential privacy allows us to inject a precisely calibrated amount of noise into the data to control the privacy risk of any calculation or statistic.

Privacy vs. Accuracy

The only way to absolutely eliminate all risk of re-identification would be to never release any usable data.

Differential privacy allows you to quantify a precise level of “acceptable risk,” and to precisely calibrate where on the privacy/accuracy spectrum the resulting data will be.


Providing accurate data
Data Quality |Bnae Kegouqe
Dada Qualitg |Vrkk Jzcfkdy
Data Qaality |Dncb PrhvBln
Dzte Qvality |Dncb Prtnavy
Dfha Quapyti |Tgta Ppijacy
Tgta Qucjity |Dfha Pnjvico
Dncb Qhulitn |Dzhe Njivaci
Ntue Quevdto |Dzte Privecy
Vrkk Zuhnvry |Dada Privacg
Bnaq Denorbe |Data Privacy
Safeguarding individual privacy

Establishing a Privacy-loss Budget

This measure is called the “Privacy-loss Budget” (PLB) or “Epsilon.”

- ε=0 (perfect privacy) would result in completely useless data
- ε=∞ (perfect accuracy) would result in releasing the data in fully identifiable form

ε

Epsilon

Comparing Methods

Data Accuracy

Differential Privacy is not inherently better or worse than traditional disclosure avoidance methods.

Both can have varying degrees of impact on data quality depending on the parameters selected and the methods’ implementation.

Privacy

Differential Privacy is substantially better than traditional methods for protecting privacy, insofar as it actually allows for measurement of the privacy risk.

Implications for the 2020 Decennial Census

The switch to Differential Privacy does not change the constitutional mandate to apportion the House of Representatives according to the actual enumeration.

As in 2000 and 2010, the Census Bureau will apply privacy protections to the PL94-171 redistricting data.

The switch to Differential Privacy requires us to re-evaluate the quantity of statistics and tabulations that we will release, because each additional statistic uses up a fraction of the privacy-loss budget (epsilon).

Demonstrating Privacy, Assessing and Improving Accuracy

The DAS Team's priorities over Fall 2019 were:

- To scale up the DAS to run on a (nearly) fully-specified national histogram
- To demonstrate that the DAS can effectively protect privacy at scale
- To permit the evaluation and optimization of the DAS for accuracy and "fitness for use"

These initiatives were largely successful, but much more work needs to be done over the remainder of this year.

The engagement and efforts of our data users have been enormously helpful in helping to identify and prioritize this remaining work.

Committee on National Statistics Workshop

December 11-12, 2019

Evaluation of the Demonstration Data Products (DDP): 2010 Census data run through a preliminary version of the 2020 DAS

Data user assessments and findings on DAS implications for:

- **Redistricting and related legal use cases**
- **Identification of rural and special populations**
- **Geospatial analysis of social/demographic conditions**
- **Delivery of government services**
- **Business and private sector applications**
- **Denominators for rates and baselines for assessments**

Shape your future START HERE >

United States Census 2020

What We’ve Learned

The October vintage of the DAS falls short on ensuring “fitness for use” for several priority use cases.

Particular areas of concern:

- Population counts for political geographies
- Population counts for American Indian and Alaska Native Tribes and Tribal Areas
- Systemic biases (e.g., urban vs. rural)
- Housing statistics and vacancy rates

These issues are substantially driven by post-processing of the noisy statistics within the DAS.

What We've Learned

- **There are two sources of error in the TopDown Algorithm (TDA):**
  - Measurement error due to differential privacy noise (tunable through selection of ε)
  - Post-processing error due to process of creating internally consistent, non-negative integer counts from the noisy measurements
- **Post-processing error tends to be much larger than DP error**
- **Improving post-processing is not constrained by DP**

Causes of Post-Processing Error

Sparsity!

Earlier runs of the DAS (e.g., 2018 E2E Test) processed a smaller histogram, where most cells were populated. (2,012 statistics = ~22 Billion cells at the block level)

The DDP included a much larger histogram. (400,000 statistics = ~4.4 Trillion cells at the block level)

The more statistics you calculate, the greater the likelihood of a pull from the tail of the noise distribution.

Within the constrained population totals of higher geographic levels of TDA, the algorithm had difficulty prioritizing legitimate positive values against all the “noisy” zeros.

Current Initiatives

Improving population totals for legal and political entities (including AIAN geograhpies)

Adopting a multi-phase approach to post-processing

- Addresses the sparsity issue
- Allows for better prioritization of use cases

Making population counts more accurate.

A set of accuracy metrics have been developed based on use cases and stakeholder feedback. The metrics will allow the public to see the improvements that are made to the Disclosure Avoidance System.

The selected metrics:

- Reflect input from external data users;
- Show differences between major DAS runs and publicly available 2010 tabulations
- Provide accuracy, bias, and outlier information for basic demographic tabulations
- Provide accuracy, bias, and outlier information for categories of use cases

These metrics will inform data users of accuracy improvements we are able to make while also informing their ongoing engagement throughout the remaining work.

Send feedback to 2020DAS@census.gov

Additional Resources

Disclosure Avoidance and the 2020 Census Website
https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html

Questions? Suggestions?
Email them to 2020DAS@census.gov

Michael Hawes
Senior Advisor for Data Access and Privacy
Research and Methodology Directorate
U.S. Census Bureau

301-763-1960 (Office)
michael.b.hawes@census.gov